OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

PATTERN MINING USING APRIORI ALGORITHM AFTER CLUSTERING OF CUSTOMER TRANSACTION DATA

Ms. Sonali Mortale[1],Prof. Mrs.Manisha Darak[2]

Department of Computer Engineering, Siddhant College of Engineering, Chakan, Pune[1,2]

sonali.mortale@gmail.com[1], darakmanisha9@gmail.com[2]

Abstract: Clustering of customer transaction data and finding patterns using Apriori Algorithm is useful task in data mining to formulate market strategies and maximize profit. The system applies an Apriori Algorithm for finding patterns. This system use output of Customer Transaction Clustering Algorithm as an input to Apriori Algorithm. The system have transaction tree which represents the customer's transaction records. Distance between transaction trees is calculated. A customer transaction clustering algorithm is used for clustering of transaction data of customers. The most frequent customers are selected as representatives of customer groups. Clustering is performed by assigning customer to the nearest neighborhood. Finally, the clustering results are forwarded to Apriori Algorithm for finding patterns.

Keywords - *Customer Transaction Clustering Algorithm, Apriori Algorithm, Transaction Tree, Clustering.*

---∴∴∴---

I INTRODUCTION

Clustering is the method of arranging objects into groups. Objects with similar characteristics are placed into one group. These groups are known as clusters. Customer Segmentation is a technique in which the customers are clustered on the basis of certain characteristics.

The main objective of this paper is to find the best number of clusters and these resultant clusters are used for finding patterns using Apriori Algorithm. In clustering, the system use large amount of raw and unorganized data as an input and determine similarities in input data.

The clustering of transaction data of customers is essential phase to identify customer activities in retail and ecommerce firms [1].As there is rapid growth in customer behavior data, scientists are now focusing on clustering of transaction data of customers [2].

Basically the transaction data is the information of the daily transactions of customers. It contains information about, what type of product or set of products purchased by customers. There are three common problems of clustering the data. One is the how to show customer and customer transaction data. Second is how to calculate the distance between different customers, and third is how to divide a customer into a certain number of customer groups.

The system apply Apriori algorithm for finding patterns. Then the system use output of Customer Transaction Clustering Algorithm as an input to Apriori Algorithm. The system have transaction tree which represents the customer's transaction records. Distance between transaction trees is calculated. A customer transaction clustering algorithm is used for clustering of transaction data of customers. The most frequent customers are selected as representatives of customer groups. Clustering is performed by assigning customer to the nearest neighborhood. Finally, the clustering results are forwarded to Apriori Algorithm for finding patterns.

Apriori algorithm is used to identify patterns. The rules are derived from the association. These rules must satisfy the minimum support threshold and the minimum confidence threshold.

The system use transaction tree distance to compare customers at all levels of the Item (Product) Tree. However, transaction data of customer are very big, even after data is compressed by transaction trees. So the speed of Clustering of customer transaction data is very important. In real applications, it is very hard to use a hierarchical clustering method because of the high computational complexity. For complex product tree data, it is very difficult to apply the fast k-means algorithm. In this paper, proposed system use clustering method called Customer Transaction Clustering for clustering of large amount of transaction tree data and uses

Apriori algorithm which further processes the output of Customer Transaction Clustering algorithm and discovers the patterns. In Customer Transaction Clustering algorithm, use a separate density for ranking a transaction trees as a representative trees. The most frequent customers are selected as representatives of customer groups and clustering is performed by assigning each customer to the nearest neighborhood. Finally, clustering results are given as an input to Apriori Algorithm for finding patterns.

The paper is organized in the following way. Literature Review is given in Section II. Proposed methodology is given in Section III. Result and discussion is given in section IV .Finally we provide conclusion in section V.

II. REVIEW OF LITERATURE

Here we present the literature review of existing techniques:

In paper [1], transaction data of customers are compressed into set of purchase trees. Distance between two purchase trees is calculated. They implemented PurTreeClust clustering algorithm for performing clustering.

In this paper [2], they considered transaction data with high capacity and dimension. The author used heuristic approach to increase the ratio of width and height of cluster histogram. They developed fast and scalable algorithm named as CLOPE.

In order to predict the yearly sales of supermarket, SPPS tool and Kmeans clustering is used to create online and real time system for supermarket [3].

SWCC Subspace Weighted Co Clustering algorithm is developed by the author for high-dimensional expression data. To identify different clusters, Subspace weight matrices were presented [4]

In this paper [5], the author presents an automatic two-level variable weighting clustering algorithm for multi view data TW-k-means, which can compute the weight of view and individual variables simultaneously. The algorithm assigns view weights to each view in order to identify the compactness of the view, assigns variable weights to each variable in the view, then using the quantifiable two real data sets, examines the nature of the two types of weight in TW-K-mean, TW-K-mean it is possible to determine the weight of the view. The difference between the weight of the man and the weight of the individual variable weighting method was examined.

The author presents a robust tree edit distance algorithm – RTED [6]. The author introduces the class of LRH (Left-Right-Heavy) algorithms, which includes RTED and the fastest tree edit distance algorithms.

In this paper [7], the author presents a new partitioning hierarchical clustering algorithm for category data named DHCC .It shows the task of clustering categorical data from an optimization perspective, and suggests an effective procedure for initializing and tuning the cluster partition. The partition initialization is based on multiple correspondence analyses (MCA). They also devise strategies to determine when to end the split process.

In this paper [8], frequent user access patterns are generated from web log entries. Combined efforts of clustering and association rule mining is used to apply pattern discovery.

In this paper, data mining techniques are used to provide customer's purchasing patterns of food items. Author uses KMedoids clustering algorithm for clustering of food items. These outputs of clustering are given as an input to the association rule mining based Apriori algorithm and frequent patterns are discovered [9].

The aim of this paper [10] is to recommend the suitable items to the user. A better Rule extraction is needed to recommend the suitable items. Association Rule mining is applied for better rule extraction .The K-means clustering algorithm method is also applied here to cluster the data based on similar characteristics.

III. PROPOSED METHODOLOGY

A. Problem Statement

To develop a system that performs clustering of customer transaction data and processes resulting clusters to Apriori Algorithm for finding patterns.

B. System Architecture

A detailed description of the proposed system is as follows:

1) Transaction Dataset:

The system uses customer transaction data. The transaction data include the products bought by customers.

2) Preprocessing of data:

Preprocessing is done on transaction data of customers.

3) Product (Item) Tree Generation:

A Product or Item tree consists number nodes. A child node represents product or item. An internal node represents category of particular item.

4) Transaction Tree Generation:

A Transaction Tree consists of number of nodes. The child nodes represent items bought by customer and internal node represents the category of particular item.

5) Transaction Tree Distance:

Customers do not buy similar products; due to this, between any two transaction trees, the tree edit distance will produce high distance value. Within the tree edit distance it is very difficult to recover the cluster structure. To solve this

issue, Transaction Tree distance metric is used. Transaction Tree distance compares customers from the entire levels of the product tree.

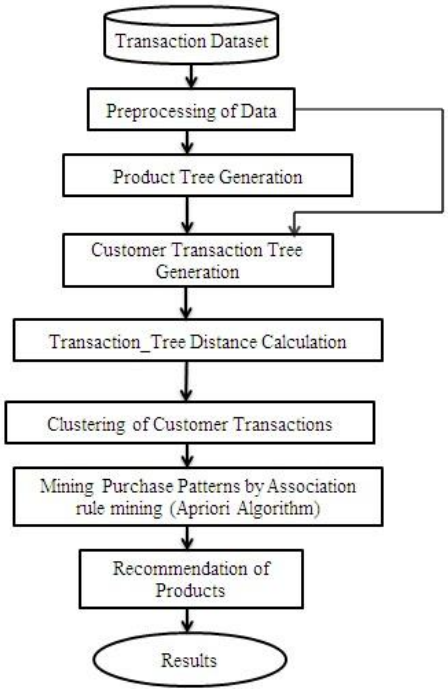

Figure 1.Proposed System Architecture

6) Transaction Tree Clustering :

Transaction Tree Clustering Algorithm is used for clustering of transaction tree data.

7) Mining Purchase patterns by association rule mining: The system use apriori algorithm for finding patterns.

8) Recommendation of products:

Finally system recommends the product and gives fast and accurate results.

C. Algorithms

- **Algorithm1: Cust_Tran_Clustering**

Clustering of customer transaction data consists of following steps.

1. Generate the Item (product) tree.
2. Generate customer transaction tree for each customer.
3. Calculate the distance between two transaction trees.
4. Estimate the level density of transaction Tree with cover tree: $den^{l}_{CT}(p)$
5. Calculate the separate distance of object $p \in CS_l$: $sdis^{l}_{CT}(p)$.
6. Calculate separate density of object $p \in CS_l$: $sden^{l}_{CT}(p)$ $sden^{l}_{CT}(p) = den^{l}_{CT}(p) * sdis^{l}_{CT}(p)$
7. Select 'k' representative trees as 'k' trees having highest separate densities.
8. Perform clustering by assigning each customer to the nearest representative.

- **Algorithm 2: Apriori Algorithm**

Following are the steps of Apriori Algorithm:

1. Initialize s=1
2. Generate frequent itemset of size '1'.
3. Generate candidate itemset of size 's+1' from frequent itemset of size 's'.
4. Prune candidate itemsets containing subsets of size 's' that are infrequent.
5. Count the support of each candidate itemset
6. Remove candidate that are infrequent, keeping only those that are frequent.
7. Repeat steps 3 to 6 until no new frequent itemsets are identified.

IV. Results And Discussion

A. Experimental Setup

For this work, required technologies:

- **Software Technology:**

1. Technology: Core Java
2. Tools: JDK 1.8, Netbeans 8.0.2
3. Operating System: Windows 7

- **Hardware Technology:**

1. System: Pentium IV 2.4 GHz
2. Hard Disc: 40 GB
3. RAM: 512 MB

B. Dataset:

To build dataset we have applied two approaches. (a) We have collected customer transaction data from offline mobile sales store. (b)We collected data from consumers through field survey campaign.

1) Transaction Data

At the end aims to understand the consumer behaviour and preferences when using mobile OSs on smart phones. The first step involves developing the database. Doing so, we have observed the trends in mobile OS industry and for this we have built the consumer survey campaign and accordingly, the tables in the relational database. The database is built on MySQL. For this we collected customer's transactions data from off-line mobile shoppers. We finalized product list of 93 products for our dataset and list of 100 customers and combine 560 transactions.

2. Survey Campaign

For more meaningful data, a field survey questionnaire approach is applied to collect information from consumers having mobile smart phones. The survey is answered by 100 consumers, where the number of male responders exceeds the number of female responders. The rate of female accounts for 44%, whereas the rate of male accounts

for 56%. Most of the respondents, 47%, are between 23-32 years old.

IV RESULTS

1) Confidence and Lift of data related to Brand:

Three rules are extracted from the demographic information, referred as R1, R2, and R3 etc. The analysis results in Table 1 reveal that Samsumg, Apple and Redmi are found more attractive among male consumers. Moreover, Redmi seems to be more desirable by the young consumers of age 23-32 years' users.

Table 1: Association Rule Result Related to Brand

Data Related to: Brand | Table | Graph | Show Results

ASSOCIATION RULES AFTER CLUSTERING RELATED TO 'BRAND'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.38	0.57	1.86	Current_Brand = {Samsung}	Next_Brand = {Apple}
R2	0.47	0.62	2.37	Current_Brand = {Samsung}	Male
R3	0.44	0.65	2.34	Current_Brand = {Apple}	Male
R4	0.49	0.69	2.36	Current_Brand = {Apple}	Female
R5	0.48	0.73	2.65	Current_Brand = {RedMi}	Male
R6	0.46	0.71	2.55	Current_Brand = {RedMi}	AGE = {23 to 32}
R7	0.39	0.55	2.15	Current_Brand = {motorola}	AGE > 40
R8	0.43	0.65	2.21	Current_Brand = {Samsung}	Female
R9	0.42	0.77	2.72	Application_Type = {Business}	Next_Brand = {Apple}
R10	0.45	0.82	2.73	Application_Type = {Social}	Next_Brand = {Samsung}

ASSOCIATION RULES BEFORE CLUSTERING RELATED TO 'BRAND'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.45	1.58	Current_Brand = {Samsung}	Next_Brand = {Apple}
R2	0.35	0.46	1.84	Current_Brand = {Samsung}	Male
R3	0.32	0.47	1.78	Current_Brand = {Apple}	Male
R4	0.39	0.53	1.96	Current_Brand = {Apple}	Female
R5	0.41	0.62	2.14	Current_Brand = {RedMi}	Male
R6	0.38	0.59	2.24	Current_Brand = {RedMi}	AGE = {23 to 32}
R7	0.27	0.49	1.76	Current_Brand = {motorola}	AGE > 40
R8	0.26	0.47	1.74	Current_Brand = {Samsung}	Female
R9	0.27	0.65	1.98	Application_Type = {Business}	Next_Brand = {Apple}
R10	0.26	0.73	2.15	Application_Type = {Social}	Next_Brand = {Samsung}

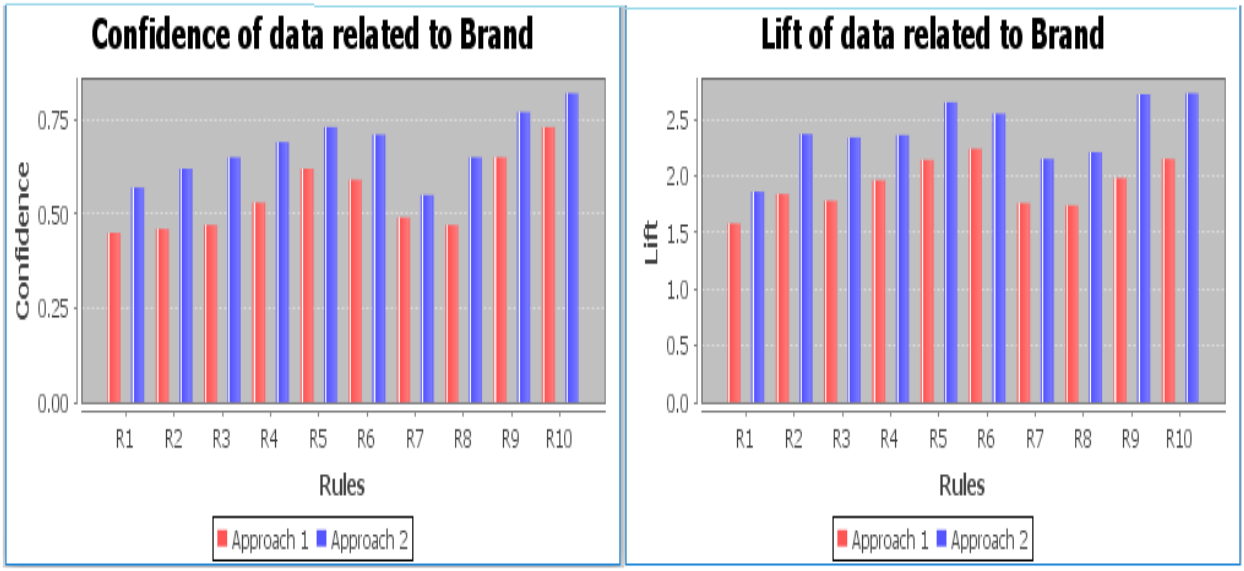

Figure 2: Comparison Between the Confidence and Lift of data related to Brand

2) Confidence and Lift of data related to OS:

The data analysis results in Table 2 show that two leaders in mobile OS market are Android, and their desirability is very close to each other. It seems that Android users want to continue to use Android as their next mobile OS. Therefore, brand loyalty has proven to be an important criterion that has influence on consumer choices, even in this small customer dataset. Telecommunication companies need to focus on attracting first time smart phone users.

Table 2: Association Rule Result Related to OS

Data Related to OS | Table | Graph | Show Results

ASSOCIATION RULES AFTER CLUSTERING RELATED TO 'OS'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.49	0.67	2.72	Next_OS = {Android}	Male
R2	0.45	0.66	2.69	Next_OS = {Android}	Female
R3	0.47	0.78	2.95	Next_OS = {Android}	AGE = {23 to 32}
R4	0.48	0.75	2.89	Current_OS = {Android}	Next_OS = {Android}
R5	0.39	0.58	2.13	Current_OS = {Android}	Next_OS = {iOS}
R6	0.41	0.74	2.03	Current_OS = {Android}	AGE >= {45}

ASSOCIATION RULES BEFORE CLUSTERING RELATED TO 'OS'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.26	0.59	2.17	Next_OS = {Android}	Male
R2	0.23	0.47	1.96	Next_OS = {Android}	Female
R3	0.31	0.62	2.37	Next_OS = {Android}	AGE = {23 to 32}
R4	0.34	0.69	2.25	Current_OS = {Android}	Next_OS = {Android}
R5	0.22	0.42	1.67	Current_OS = {Android}	Next_OS = {iOS}
R6	0.29	0.69	1.72	Current_OS = {Android}	AGE >= {45}

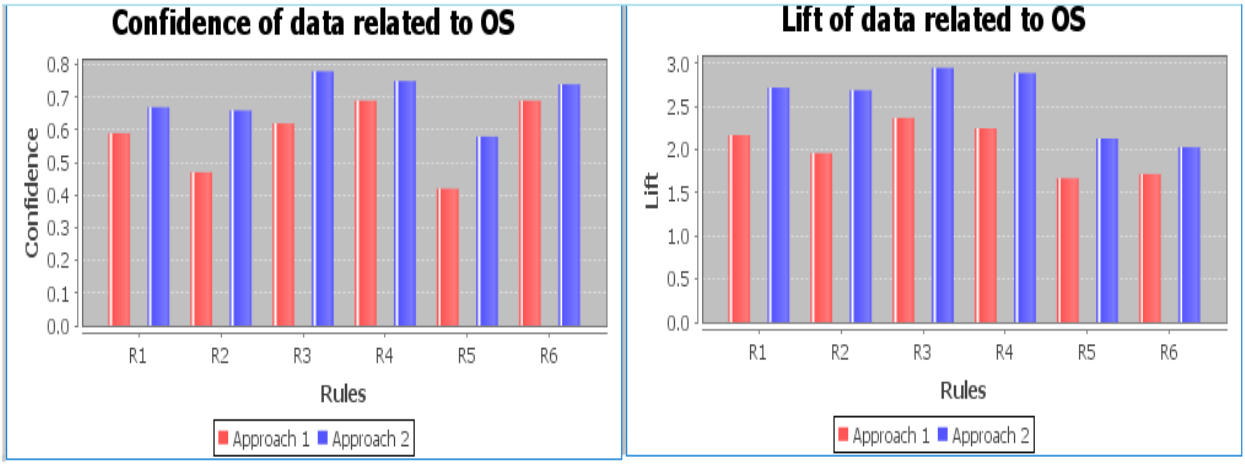

Figure 3. Comparison between the Confidence and Lift of data related to OS

Result graph shows that the proposed system is more accurate than the existing system.

V. CONCLUSION

The system uses Apriori Algorithm for finding patterns. Then the system use output of Customer Transaction Clustering Algorithm as an input to Apriori Algorithm. Apriori Algorithm is used for finding patterns. This is beneficial to increase product sale by identifying relations between combinations of customer's purchase pattern.

A customer transaction clustering algorithm is used for clustering of transaction data of customers. The most frequent customers are selected as representatives of customer groups. Clustering is performed by assigning customer to the nearest neighborhood. Finally, the clustering results are forwarded to Apriori Algorithm for finding patterns.

From the graph, conclude that Apriori algorithm with Customer Transaction Clustering is more accurate and efficient than Apriori Algorithm without Clustering.

Resultant patterns are useful to formulate market strategies and maximize profit.

REFERENCES

[1] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data," IEEE vol. 30, no. 3, March 2018.

[2] Y. Yang ,X. Guan,J. You , "CLOPE: A Fast and Effective Clustering Algorithm for Transaction Data" in Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mmining ,ACM 2002

[3] Kishana R. Kashwan and C.M.Velu , "Customer Segmentation Using Clustering and Data Mining Techniques",IJCTE Vol. 5.No.6 December 2013

[4] Q. Wu, X. Chen ,J. Z. Huang and M. Yang , "Subspace Weighting Co_clustering of Gene Expression Data", TCBB, 2017.

[5] X. Chen, X. Xu,Y. Ye and J.Z. Huang , "TW-k-means: Automated two level variable weighting clustering algorithm for multi-view data", IEEE Vol.25 No.4 Apr. 2013

[6] M. Pawlik and N. Augsten , "RTED: A robust algorithm for the tree edit distance",Proc.VLDB Endowment ,Vol.5,No.4 ,2011

[7] T.Xiong, S.Wang, A.Mayersand E.Monga "DHCC: Divisive Hierarchical Clustering of Categorical Data", Data Mining Knowledge Discovery, Vol.24, No.1, 2012.

[8] Htun Zaw Oo, Nang Saing Moon Kham, "Pattern Discovery Using Association Rule Mining on Clustered Data",IJNTR,ISSN:2454-4116,Volume-4, Issue-2,February 2018,Page 07-11.

[9] Kavita M. Gawande .Mr. Subhash K. Shinde , Mrs. Dipti Patil , "Frequent Pattern Mining Based on Clustering and Association Rule Algorithm", IJARCS,Vol.3,No.3,2012.

[10] Jaimeel. M. Shah Lokesh Sahu, "Recommendation based on Clustering and Association Rules", IJARIIE-ISSN (O)-2395-4396, Vol-1 Issue-2 2015.